第十章

极大似然估计

Outline

Maximum Likelihood Estimation

The basic idea of the ML principle is to choose the parameter estimates to maximize the probability of obtaining the data.

定义 (Model)

A model for a random sample is the assumption that X_i, $i = 1, \ldots, n$, are i.i.d. with known density function $f(x|\boldsymbol{\theta})$ or mass function $\pi(x|\boldsymbol{\theta})$ with unknown parameter $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^p$.

定义 (Correctly Specified Model)

A model is correctly specified when there is a unique parameter value $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}$ such that $f(x|\boldsymbol{\theta}_0) = f(x)$, the true data distribution. This parameter value $\boldsymbol{\theta}_0$ is called the true parameter value. The parameter $\boldsymbol{\theta}_0$ is unique if there is no other $\boldsymbol{\theta}$ such that $f(x|\boldsymbol{\theta}_0) = f(x|\boldsymbol{\theta})$. A model is mis-specified if there is no parameter value $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ such that $f(x|\boldsymbol{\theta}) = f(x)$.

Likelihood

The joint density evaluated at the observed data $(x_1, x_2, ..., x_n)$ and viewed as a function of $\boldsymbol{\theta}$ is called the likelihood function:

$$L_n(\boldsymbol{\theta}) \equiv f\left(x_1, \ldots, x_n | \boldsymbol{\theta}\right) = \prod_{i=1}^{n} f\left(x_i | \boldsymbol{\theta}\right)$$

for continuous random variables and

$$L_n(\boldsymbol{\theta}) \equiv \prod_{i=1}^{n} \pi\left(x_i | \boldsymbol{\theta}\right)$$

for discrete random variables.

The value of $\boldsymbol{\theta}$ most compatible with the observations is the value which maximizes the likelihood.

定义 (MLE)

The maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ is the value which maximizes $L_n(\boldsymbol{\theta})$:

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\operatorname{argmax}} \, L_n(\boldsymbol{\theta}). \tag{1}$$

In most cases it is more convenient to calculate and maximize the logarithm of the likelihood function rather than the level of the likelihood.

定义 (Log-likelihood)

The log-likelihood function is

$$\ell_n(\boldsymbol{\theta}) \equiv \log L_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f\left(x_i | \boldsymbol{\theta}\right).$$

The maximizer of the likelihood and log-likelihood function are the same, since the logarithm is a monotonically increasing function.

Define the expected log density function

$$\ell(\boldsymbol{\theta}) = \mathbb{E}[\log f(X|\boldsymbol{\theta})].$$

Theorem

When the model is correctly specified the true parameter $\boldsymbol{\theta}_0$ maximizes the expected log density $\ell(\boldsymbol{\theta})$:

$$\boldsymbol{\theta}_0 = \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\operatorname{argmax}}\, \ell(\boldsymbol{\theta}).$$

The sample analog of $\ell(\boldsymbol{\theta})$ is the average log-likelihood

$$\bar{\ell}_n(\boldsymbol{\theta}) = \frac{1}{n}\ell_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \log f\left(x_i|\boldsymbol{\theta}\right)$$

which has maximizer $\hat{\boldsymbol{\theta}}$.

证明.

Take any $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$. Since the difference of logs is the log of the ratio

$$
\begin{aligned}
\ell(\boldsymbol{\theta}) - \ell(\boldsymbol{\theta}_0) &= \mathbb{E}\left[\log(f(X|\boldsymbol{\theta})) - \log(f(X|\boldsymbol{\theta}_0))\right] \\
&= \mathbb{E}\left[\log\left(\frac{f(X|\boldsymbol{\theta})}{f(X|\boldsymbol{\theta}_0)}\right)\right] < \log\left(\mathbb{E}\left[\frac{f(X|\boldsymbol{\theta})}{f(X|\boldsymbol{\theta}_0)}\right]\right),
\end{aligned}
$$

where the last inequality follows from Jensen's inequality and is strict since $f(x|\boldsymbol{\theta}) \neq f(x|\boldsymbol{\theta}_0)$ with positive probability. Note that

$$
\begin{aligned}
\log\left(\int \frac{f(x|\boldsymbol{\theta})}{f(x|\boldsymbol{\theta}_0)} f(x) dx\right) &= \log\left(\int \frac{f(x|\boldsymbol{\theta})}{f(x|\boldsymbol{\theta}_0)} f(x|\boldsymbol{\theta}_0)\, dx\right) \\
&= \log\left(\int f(x|\boldsymbol{\theta}) dx\right) = 0.
\end{aligned}
$$

Therefore, $\ell(\boldsymbol{\theta}) < \ell(\boldsymbol{\theta}_0)$. □

Invariance Property

A special property of the MLE is that it is invariant to transformations.

Theorem

If $\hat{\boldsymbol{\theta}}$ is the MLE of $\boldsymbol{\theta}$ then for any transformation $\boldsymbol{\beta} = \boldsymbol{h}(\boldsymbol{\theta})$ the MLE of $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}} = \boldsymbol{h}(\hat{\boldsymbol{\theta}})$.

证明.

We can write the likelihood for the transformed parameter as

$$L_n^*(\boldsymbol{\beta}) = \max_{\boldsymbol{h}(\boldsymbol{\theta})=\boldsymbol{\beta}} L_n(\boldsymbol{\theta}).$$

The MLE for $\boldsymbol{\beta}$ maximizes $L_n^*(\boldsymbol{\beta})$. Evaluating $L_n^*(\boldsymbol{\beta})$ at $\boldsymbol{h}(\hat{\boldsymbol{\theta}})$ we find

$$L_n^*(\boldsymbol{h}(\hat{\boldsymbol{\theta}})) = \max_{\boldsymbol{h}(\boldsymbol{\theta})=\boldsymbol{h}(\hat{\boldsymbol{\theta}})} L_n(\boldsymbol{\theta}) = L_n(\hat{\boldsymbol{\theta}}).$$

The final equality holds because $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$ satisfies $\boldsymbol{h}(\boldsymbol{\theta}) = \boldsymbol{h}(\hat{\boldsymbol{\theta}})$ and maximizes $L_n(\boldsymbol{\theta})$. □

To find the MLE we take the following steps:

1. Construct $f(x|\boldsymbol{\theta})$ as a function of x and $\boldsymbol{\theta}$
2. Take the logarithm $\log f(x|\boldsymbol{\theta})$
3. Evaluate at $x = x_i$ and sum over i : $\ell_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f\left(x_i|\boldsymbol{\theta}\right)$
4. If possible, solve the first order condition (F.O.C.) to find the maximum
5. If solving the F.O.C. is not possible use numerical methods to maximize $\ell_n(\boldsymbol{\theta})$

Conditional ML Estimation of the Classical Linear Regression Model

Conditional Maximum Likelihood

In most applications, observation for i is partitioned into two groups, y_i and $\mathbf{x}_i$. Let $f(y_i|\mathbf{x}_i;\boldsymbol{\theta}_0)$ be the conditional density of y_i, given $\mathbf{x}_i$, and let $f(\mathbf{x}_i;\boldsymbol{\psi}_0)$ be the marginal density of $\mathbf{x}_i$. Then

$$f(y_i,\mathbf{x}_i;\boldsymbol{\theta}_0,\boldsymbol{\psi}_0)=f(y_i|\mathbf{x}_i;\boldsymbol{\theta}_0)f(\mathbf{x}_i;\boldsymbol{\psi}_0).$$

Let $\mathbf{w}_i=(y_i,\mathbf{x}_i)$. The log likelihood is

$$\sum_{i=1}^{n}\log f(\mathbf{w}_i;\boldsymbol{\theta},\boldsymbol{\psi})=\sum_{i=1}^{n}\log f(y_i|\mathbf{x}_i;\boldsymbol{\theta})+\sum_{i=1}^{n}\log f(\mathbf{x}_i;\boldsymbol{\psi}).$$

The first term on the right-hand side is the log conditional likelihood. The conditional ML estimator of $\boldsymbol{\theta}_0$ maximizes this first term. If the second term does not depend on $\boldsymbol{\theta}$, the conditional ML estimator of $\boldsymbol{\theta}_0$ is numerically the same as the joint ML estimator.

Under the assumptions of classical linear regression, we have

$$\begin{aligned}\mathbf{y} &= \mathbf{X}\boldsymbol{\beta}_0 + \boldsymbol{\varepsilon} \\ \mathbf{y} \mid \mathbf{X} &\sim N(\mathbf{X}\boldsymbol{\beta}_0, \sigma_0^2 \mathbf{I}_n).\end{aligned}$$

Thus, the conditional density of $\mathbf{y}$ given $\mathbf{X}$ is[1]

$$f(\mathbf{y}|\mathbf{X}; \boldsymbol{\beta}_0, \sigma_0^2) = (2\pi\sigma_0^2)^{-n/2} \exp\left[-\frac{1}{2\sigma_0^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_0)'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_0)\right].$$

Therefore, the log likelihood function is

$$\log L(\boldsymbol{\beta}, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).$$

[1] The density function for an n-variate normal distribution $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is

$$(2\pi)^{-n/2}|\boldsymbol{\Sigma}|^{-1/2} \exp\left[-\frac{1}{2}(\mathbf{y} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{y} - \boldsymbol{\mu})\right].$$

ML via Concentrated Likelihood

The log likelihood function can be maximized in two stages:

1. maximize over $\boldsymbol{\beta}$ for any given σ^2;
2. maximize over σ^2 taking into account that the $\boldsymbol{\beta}$ obtained in the first stage could depend on σ^2.

The log likelihood function in which $\boldsymbol{\beta}$ is constrained to be the value from the first stage is called the concentrated log likelihood function. It is easy to see that

$$\text{concentrated log likelihood} = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\boldsymbol{e}'\boldsymbol{e}.$$

定理 (ML Estimator of $(\boldsymbol{\beta}, \sigma^2)$)

Suppose Assumptions 1.1–1.5 hold. Then $\hat{\boldsymbol{\beta}}_{MLE} = \boldsymbol{b}$ *and* $\hat{\sigma}^2_{MLE} = \boldsymbol{e}'\boldsymbol{e}/n$.

Score, Hessian, and Information

The likelihood score is the derivative of the likelihood function:

$$\boldsymbol{S}_n(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} \ell_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\theta}} \log f\left(x_i | \boldsymbol{\theta}\right).$$

Note that $\boldsymbol{S}_n(\hat{\boldsymbol{\theta}}) = \mathbf{0}$ when $\hat{\boldsymbol{\theta}}$ is an interior solution.

The likelihood Hessian is the negative second derivative of the likelihood function:

$$\mathscr{H}_n(\boldsymbol{\theta}) = -\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \ell_n(\boldsymbol{\theta}) = -\sum_{i=1}^{n} \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \log f\left(x_i | \boldsymbol{\theta}\right).$$

The efficient score is the derivative of the log-likelihood for a single observation, evaluated at the random variable X and the true parameter vector:

$$\boldsymbol{S} = \frac{\partial}{\partial \boldsymbol{\theta}} \log f\left(X | \boldsymbol{\theta}_0\right).$$

Theorem

Assume that the model is correctly specified, the support of X does not depend on $\boldsymbol{\theta}$, and $\boldsymbol{\theta}_0$ lies in the interior of $\boldsymbol{\Theta}$. Then the efficient score $\boldsymbol{S}$ satisfies $\mathbb{E}[\boldsymbol{S}] = \mathbf{0}$.

证明.

$$\mathbb{E}[\boldsymbol{S}] = \mathbb{E}\left[\frac{\partial}{\partial \boldsymbol{\theta}} \log f\left(X|\boldsymbol{\theta}_0\right)\right] = \frac{\partial}{\partial \boldsymbol{\theta}} \mathbb{E}\left[\log f\left(X|\boldsymbol{\theta}_0\right)\right] = \frac{\partial}{\partial \boldsymbol{\theta}} \ell\left(\boldsymbol{\theta}_0\right) = \mathbf{0}.$$

□

- The assumption that the support does not depend on the parameter is needed for interchange of integration and differentiation.
- The assumption that $\boldsymbol{\theta}_0$ lies in the interior of $\boldsymbol{\Theta}$ is needed to ensure that the expected log-likelihood $\ell(\boldsymbol{\theta})$ satisfies a first-order-condition.

The Fisher information is the variance of the efficient score:

$$\mathscr{I}_{\boldsymbol{\theta}} = \mathbb{E}\left[\boldsymbol{S}\boldsymbol{S}'\right].$$

The expected Hessian is

$$\mathscr{H}_{\boldsymbol{\theta}} = -\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\ell\left(\boldsymbol{\theta}_0\right).$$

Under regularity conditions,

$$\mathscr{H}_{\boldsymbol{\theta}} = -\mathbb{E}\left[\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\log f\left(X|\boldsymbol{\theta}_0\right)\right].$$

Theorem: Information Matrix Equality

Assume that the model is correctly specified and the support of X does not depend on $\boldsymbol{\theta}$. Then the Fisher information equals the expected Hessian:

$$\mathscr{I}_{\boldsymbol{\theta}} = \mathscr{H}_{\boldsymbol{\theta}}.$$

证明.

The expected Hessian equals

$$\begin{aligned}
\mathscr{H}_{\boldsymbol{\theta}} &= -\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\mathbb{E}\left[\log f\left(X|\boldsymbol{\theta}_0\right)\right] = -\frac{\partial}{\partial\boldsymbol{\theta}}\mathbb{E}\left[\frac{\frac{\partial}{\partial\boldsymbol{\theta}'}f(X|\boldsymbol{\theta})}{f(X|\boldsymbol{\theta})}\right]\Bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \\
&= \underbrace{-\frac{\partial}{\partial\boldsymbol{\theta}}\mathbb{E}\left[\frac{\frac{\partial}{\partial\boldsymbol{\theta}'}f(X|\boldsymbol{\theta})}{f\left(X|\boldsymbol{\theta}_0\right)}\right]\Bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}}_{\mathbf{0}} + \underbrace{\mathbb{E}\left[\frac{\frac{\partial}{\partial\boldsymbol{\theta}}f\left(X|\boldsymbol{\theta}_0\right)\frac{\partial}{\partial\boldsymbol{\theta}'}f\left(X|\boldsymbol{\theta}_0\right)}{f\left(X|\boldsymbol{\theta}_0\right)^2}\right]}_{\mathscr{I}_{\boldsymbol{\theta}}}.
\end{aligned}$$

$\square$

Cramér-Rao Lower Bound

The information matrix provides a lower bound for the variance among unbiased estimators.

Theorem: Cramér-Rao Lower Bound

Assume that the model is correctly specified, the support of X does not depend on $\boldsymbol{\theta}$, and $\boldsymbol{\theta}_0$ lies in the interior of $\boldsymbol{\Theta}$. If $\tilde{\boldsymbol{\theta}}$ is an unbiased estimator of $\boldsymbol{\theta}$ then $\mathrm{var}[\tilde{\boldsymbol{\theta}}] \geq (n\mathscr{I}_{\boldsymbol{\theta}})^{-1}$.

- The Cramér-Rao Lower Bound (CRLB) is $(n\mathscr{I}_{\boldsymbol{\theta}})^{-1}$.
- An estimator $\tilde{\boldsymbol{\theta}}$ is Cramér-Rao efficient if it is unbiased for $\boldsymbol{\theta}$ and $\mathrm{var}[\tilde{\boldsymbol{\theta}}] = (n\mathscr{I}_{\boldsymbol{\theta}})^{-1}$.

证明.

Write $\boldsymbol{x} = (x_1, \ldots, x_n)'$ and $\boldsymbol{X} = (X_1, \ldots, X_n)'$. The joint density of $\boldsymbol{X}$ is $f(\boldsymbol{x}|\boldsymbol{\theta}) = f(x_1|\boldsymbol{\theta}) \cdots f(x_n|\boldsymbol{\theta})$. The likelihood score is $\boldsymbol{S}_n(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} \log f(\boldsymbol{x}|\boldsymbol{\theta})$. Unbiasedness of $\tilde{\boldsymbol{\theta}}$ means that, for all $\boldsymbol{\theta}$,

$$\boldsymbol{\theta} = \mathbb{E}_{\boldsymbol{\theta}}[\tilde{\boldsymbol{\theta}}(\boldsymbol{X})] = \int \tilde{\boldsymbol{\theta}}(\boldsymbol{x}) f(\boldsymbol{x}|\boldsymbol{\theta}) d\boldsymbol{x}.$$

The vector derivative of the left side is

$$\frac{\partial}{\partial \boldsymbol{\theta}'} \boldsymbol{\theta} = \boldsymbol{I}_p.$$

and the vector derivative of the right side is

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{\theta}'} \int \tilde{\boldsymbol{\theta}}(\boldsymbol{x}) f(\boldsymbol{x}|\boldsymbol{\theta}) d\boldsymbol{x} &= \int \tilde{\boldsymbol{\theta}}(\boldsymbol{x}) \frac{\partial}{\partial \boldsymbol{\theta}'} f(\boldsymbol{x}|\boldsymbol{\theta}) d\boldsymbol{x} \\ &= \int \tilde{\boldsymbol{\theta}}(\boldsymbol{x}) \frac{\partial}{\partial \boldsymbol{\theta}'} \log f(\boldsymbol{x}|\boldsymbol{\theta}) f(\boldsymbol{x}|\boldsymbol{\theta}) d\boldsymbol{x}. \end{aligned}$$

证明.

Evaluated at the true value this is

$$\mathbb{E}\left[\tilde{\boldsymbol{\theta}}(\boldsymbol{X})\boldsymbol{S}_n(\boldsymbol{\theta}_0)'\right] = \mathbb{E}\left[\left(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)\boldsymbol{S}_n(\boldsymbol{\theta}_0)'\right], \quad \text{since } \mathbb{E}\left[\boldsymbol{S}_n(\boldsymbol{\theta}_0)\right] = \mathbf{0}.$$

Therefore, the variance matrix of stacked $\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0$ and $\boldsymbol{S}_n(\boldsymbol{\theta}_0)$ is $\begin{bmatrix} \boldsymbol{V} & \boldsymbol{I}_p \\ \boldsymbol{I}_p & n\mathscr{I}_{\boldsymbol{\theta}} \end{bmatrix}$, where $\boldsymbol{V} = \operatorname{var}[\tilde{\boldsymbol{\theta}}]$. Since it is a variance matrix it is positive semi-definite and thus remains so if we pre- and post-multiply by the same matrix, e.g.

$$\begin{bmatrix} \boldsymbol{I}_p & -(n\mathscr{I}_{\boldsymbol{\theta}})^{-1} \end{bmatrix}\begin{bmatrix} \boldsymbol{V} & \boldsymbol{I}_p \\ \boldsymbol{I}_p & n\mathscr{I}_{\boldsymbol{\theta}} \end{bmatrix}\begin{bmatrix} \boldsymbol{I}_p \\ -(n\mathscr{I}_{\boldsymbol{\theta}})^{-1} \end{bmatrix} = \boldsymbol{V} - (n\mathscr{I}_{\boldsymbol{\theta}})^{-1} \geq \mathbf{0}.$$

□

Cramér-Rao Lower Bound for $\boldsymbol{b}$

定理 ($\boldsymbol{b}$ is the Best Unbiased Estimator (BUE))

Under Assumptions 1.1–1.5, the OLS estimator $\mathbf{b}$ *is BUE in that any other unbiased (but not necessarily linear) estimator has larger conditional variance in the matrix sense.*

The ML estimator of σ_0^2 is biased, so the Cramér-Rao bound does not apply.

Consistency and Asymptotic Normality of MLE

Uniform Convergence in Probability

定义

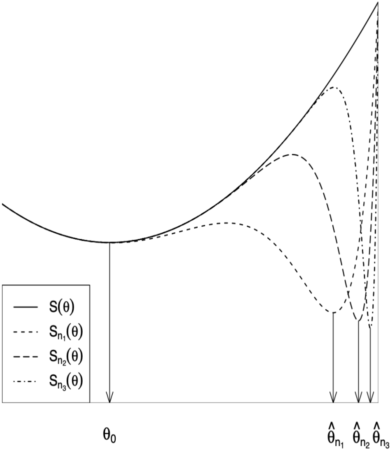
Pointwise convergence in probability of $\bar{\ell}_n(\cdot)$ to $\ell(\cdot)$ means that the sequence of random variables $|\bar{\ell}_n(\boldsymbol{\theta}) - \ell(\boldsymbol{\theta})|$ $(n = 1, 2, \ldots)$ converges in probability to 0 for each $\boldsymbol{\theta}$. Uniform convergence in probability means

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} |\bar{\ell}_n(\boldsymbol{\theta}) - \ell(\boldsymbol{\theta})| \underset{p}{\to} 0 \quad \text{as} \quad n \to \infty.$$

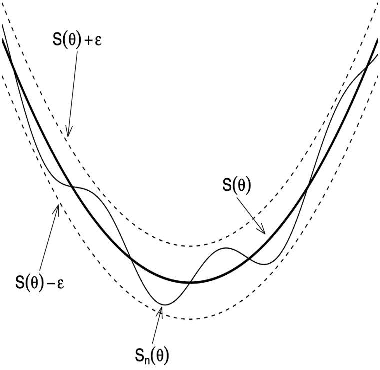
A sequence of vector random functions $\{\mathbf{h}_n(\cdot)\}$ converges uniformly in probability to a nonrandom function $\mathbf{h}(\cdot)$ if

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \|\mathbf{h}_n(\boldsymbol{\theta}) - \mathbf{h}(\boldsymbol{\theta})\| \underset{p}{\to} 0 \quad \text{as} \quad n \to \infty,$$

where $\|\cdot\|$ is the usual Euclidean norm.



(a) Non-Uniform Convergence



(b) Uniform Convergence

Existence and Consistency

定理 (Existence of MLE)

Suppose that (i) the parameter space $\boldsymbol{\Theta}$ *is a compact subset of* $\mathbb{R}^p$*, (ii)* $\log f\left(x_i|\boldsymbol{\theta}\right)$ *is continuous in* $\boldsymbol{\theta}$ *for each* x_i*, and (iii)* $\log f\left(x_i|\boldsymbol{\theta}\right)$ *is a measurable in* x_i *for all* $\boldsymbol{\theta}$ *in* $\boldsymbol{\Theta}$*. Then there exists a measurable function* $\hat{\boldsymbol{\theta}}$ *of the data that solves (1).*

定理 (Consistency of MLE)

Let $\{X_i\}$ *be i.i.d. Suppose that the conditions for Existence hold. Suppose, further, that*

(a) *(identification)* $\ell(\boldsymbol{\theta})$ *is uniquely maximized on* $\boldsymbol{\Theta}$ *at* $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}$.

(b) *(uniform convergence)* $\bar{\ell}_n(\boldsymbol{\theta})$ *converges uniformly in probability to* $\ell(\boldsymbol{\theta})$,

then $\hat{\boldsymbol{\theta}} \underset{p}{\rightarrow} \boldsymbol{\theta}_0$.

证明.

Let N be an open neighborhood in $\mathbb{R}^p$ containing $\boldsymbol{\theta}_0$. Then $N^C \cap \boldsymbol{\Theta}$ is compact. Therefore $\max_{\boldsymbol{\theta}\in N^C\cap\boldsymbol{\Theta}} \ell(\boldsymbol{\theta})$ exists. Denote

$$\epsilon = \ell(\boldsymbol{\theta}_0) - \max_{\boldsymbol{\theta}\in N^C\cap\boldsymbol{\Theta}} \ell(\boldsymbol{\theta}). \tag{2}$$

Let A_n be the event "$|\bar{\ell}_n(\boldsymbol{\theta}) - \ell(\boldsymbol{\theta})| < \epsilon/2$ for all $\boldsymbol{\theta}$." Then

$$A_n \quad \Rightarrow \quad \ell(\hat{\boldsymbol{\theta}}) > \bar{\ell}_n(\hat{\boldsymbol{\theta}}) - \epsilon/2 \tag{3}$$

and

$$A_n \quad \Rightarrow \quad \bar{\ell}_n(\boldsymbol{\theta}_0) > \ell(\boldsymbol{\theta}_0) - \epsilon/2. \tag{4}$$

证明.

But, since $\bar{\ell}_n(\hat{\boldsymbol{\theta}}) \geq \bar{\ell}_n(\boldsymbol{\theta}_0)$, we have from (3)

$$A_n \quad \Rightarrow \quad \ell(\hat{\boldsymbol{\theta}}) > \bar{\ell}_n(\boldsymbol{\theta}_0) - \epsilon/2. \tag{5}$$

Therefore, adding both sides of (4) and (5), we obtain

$$A_n \quad \Rightarrow \quad \ell(\hat{\boldsymbol{\theta}}) > \ell(\boldsymbol{\theta}_0) - \epsilon. \tag{6}$$

Therefore, from (2) and (6) we can conclude

$$A_n \quad \Rightarrow \quad \hat{\boldsymbol{\theta}} \in N,$$

which implies

$$\mathrm{Prob}(A_n) \leq \mathrm{Prob}(\hat{\boldsymbol{\theta}} \in N).$$

But, since $\lim\limits_{n\to\infty} \mathrm{Prob}(A_n) = 1$ by (b), $\hat{\boldsymbol{\theta}} \underset{p}{\rightarrow} \boldsymbol{\theta}_0$. □

Asymptotic Normality

定理 (Asymptotic Normality of MLE)

Let $\{X_i\}$ be i.i.d. Suppose that the conditions for $\hat{\boldsymbol{\theta}}$ to be consistent are satisfied. Suppose, further, that

(1) $\boldsymbol{\theta}_0$ *is in the interior of* $\boldsymbol{\Theta}$,

(2) $f(x_i|\boldsymbol{\theta})$ *is twice continuously differentiable in* $\boldsymbol{\theta}$ *for any* x_i,

(3) $\mathbb{E}[\frac{\partial}{\partial \boldsymbol{\theta}} \log f\left(X_i|\boldsymbol{\theta}_0\right)] = \mathbf{0}$ *and* $\mathscr{H}_{\boldsymbol{\theta}} = \mathscr{I}_{\boldsymbol{\theta}}$,

(4) *for some neighborhood* $\boldsymbol{\mathcal{N}}$ *of* $\boldsymbol{\theta}_0$, $\mathbb{E}\left[\sup_{\boldsymbol{\theta} \in \boldsymbol{\mathcal{N}}} \left\| \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \log f\left(X_i|\boldsymbol{\theta}\right) \right\| \right] < \infty$,

(5) $\mathscr{H}_{\boldsymbol{\theta}}$ *is nonsingular.*

Then $\hat{\boldsymbol{\theta}}$ *is asymptotically normal with*

$$\mathrm{Avar}(\hat{\boldsymbol{\theta}}) = \mathscr{I}_{\boldsymbol{\theta}}^{-1}.$$

证明.

By Mean Value Theorem, we have

$$\begin{aligned}\frac{\partial \bar{\ell}_n(\hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}} &= \frac{\partial \bar{\ell}_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} + \frac{\partial^2 \bar{\ell}_n(\bar{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \\ &= \frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\theta}} \log f\left(x_i|\boldsymbol{\theta}_0\right) + \left[\frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \log f\left(x_i|\bar{\boldsymbol{\theta}}\right)\right](\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \\ &= \mathbf{0},\end{aligned}$$

where $\bar{\boldsymbol{\theta}}$ lies between $\boldsymbol{\theta}_0$ and $\hat{\boldsymbol{\theta}}$. Therefore,

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = -\left[\frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \log f\left(x_i|\bar{\boldsymbol{\theta}}\right)\right]^{-1} \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\theta}} \log f\left(x_i|\boldsymbol{\theta}_0\right).$$

□

证明.

By i.i.d. and assumption (4), we have

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial}{\partial\boldsymbol{\theta}}\log f\left(x_i|\boldsymbol{\theta}_0\right) \underset{d}{\rightarrow} N(\mathbf{0},\mathscr{I}_{\boldsymbol{\theta}}),$$

$$-\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\log f\left(x_i|\bar{\boldsymbol{\theta}}\right) \underset{p}{\rightarrow} \mathscr{H}_{\boldsymbol{\theta}}.$$

Therefore,

$$\sqrt{n}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)\underset{d}{\rightarrow}\mathscr{H}_{\boldsymbol{\theta}}^{-1}N(\mathbf{0},\mathscr{I}_{\boldsymbol{\theta}})=N(\mathbf{0},\mathscr{I}_{\boldsymbol{\theta}}^{-1}).$$

□

$\mathrm{Avar}(\hat{\boldsymbol{\theta}})$ can be consistently estimated by

- first estimator of $\mathrm{Avar}(\hat{\boldsymbol{\theta}})$:

$$-\left\{\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\log f(x_i|\hat{\boldsymbol{\theta}})\right\}^{-1}.$$

- second estimator of $\mathrm{Avar}(\hat{\boldsymbol{\theta}})$:

$$\left\{\frac{1}{n}\sum_{i=1}^{n}\left[\frac{\partial}{\partial\boldsymbol{\theta}}\log f(x_i|\hat{\boldsymbol{\theta}})\right]\left[\frac{\partial}{\partial\boldsymbol{\theta}'}\log f(x_i|\hat{\boldsymbol{\theta}})\right]\right\}^{-1}.$$

Wald, LM, and LR Test

The Null Hypothesis

Let $\boldsymbol{\theta}_0$ be the p-dimensional model parameter. The null hypothesis can be expressed as

$$\mathbb{H}_0 : \underset{(r\times 1)}{\mathbf{c}(\boldsymbol{\theta}_0)} = \mathbf{0}.$$

We assume that $\mathbf{c}(\cdot)$ is continuously differentiable. Also, let

$$\underset{(r\times p)}{\mathbf{C}(\boldsymbol{\theta})} \equiv \frac{\partial \mathbf{c}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}.$$

We assume that

$$\underset{(r\times p)}{\mathbf{C}_0} \equiv \mathbf{C}(\boldsymbol{\theta}_0) \text{ is of full row rank.}$$

Let $\hat{\boldsymbol{\theta}}$ solve the unconstrained optimization problem. The constrained estimator, denoted $\tilde{\boldsymbol{\theta}}$, solves

$$\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} L_n(\boldsymbol{\theta}) \quad \text{s.t.} \quad \mathbf{c}(\boldsymbol{\theta}) = \mathbf{0}.$$

- Wald Statistic

$$W \equiv n\mathbf{c}(\hat{\boldsymbol{\theta}})' \left[\mathbf{C}(\hat{\boldsymbol{\theta}})\widehat{\mathscr{I}_{\boldsymbol{\theta}}}^{-1}\mathbf{C}(\hat{\boldsymbol{\theta}})'\right]^{-1} \mathbf{c}(\hat{\boldsymbol{\theta}}) \underset{d}{\rightarrow} \chi^2(r)$$

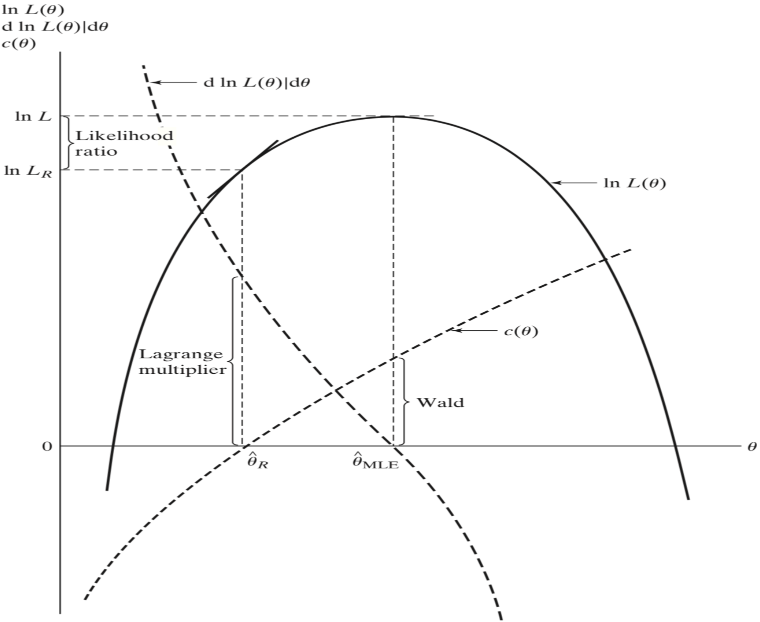
- Lagrange Multiplier (LM) Statistic (or, Score Test Statistic)

$$LM \equiv n \left(\frac{\partial \bar{\ell}_n(\tilde{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}}\right)' \widetilde{\mathscr{I}_{\boldsymbol{\theta}}}^{-1} \left(\frac{\partial \bar{\ell}_n(\tilde{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}}\right) \underset{d}{\rightarrow} \chi^2(r)$$

- Likelihood Ratio (LR) Statistic

$$LR \equiv 2n \cdot [\bar{\ell}_n(\hat{\boldsymbol{\theta}}) - \bar{\ell}_n(\tilde{\boldsymbol{\theta}})] \underset{d}{\rightarrow} \chi^2(r)$$

To summarize, the three statistics all converge in distribution to $\chi^2(r)$ under the null where r is the number of restrictions in the null.



$\ln L(\theta)$
$d \ln L(\theta)|d\theta$
$c(\theta)$
$d \ln L(\theta)|d\theta$
$\ln L$
Likelihood ratio
$\ln L_R$
$\ln L(\theta)$
$c(\theta)$
Lagrange multiplier
Wald
0
$\hat{\theta}_R$
$\hat{\theta}_{\text{MLE}}$
θ

References

Fumio Hayashi. (2000)
Econometrics, Chapter 1.
Princeton University Press, 2000.

Bruce E. Hansen. (2020)
Introduction to Econometrics, Chapter 10.
https://www.ssc.wisc.edu/ bhansen/econometrics/ Link